<Contributed Chair> Proteome Informatics (SGI Japan)

Contributed Chair
- Proteome Informatics (SGI Japan) -

http://www.bic.kyoto-u.ac.jp/proteome/index.html

Vis Assoc Prof
MAMITSUKA, Hiroshi
(D Sc)

Vis Instr
YAMAGUCHI, Atsuko

PD
AOKI, Kiyoko F.
(Ph D)

PD
MAJEUX, Nicolas
(Ph D)

Scope of Research

The objective of this laboratory is to establish and develop computationally efficient methods and algorithms utilizing vast amounts of data accumulated from genomics and proteomics to better understand biologically important phenomena. These methods and algorithms form the foundation for analytical systems useful for biology and related fields, such as (bio)chemistry, pharmacology and medical science. Our research themes focus on issues related to proteins, with particular emphasis on protein-protein and protein-ligand interactions.

Research Activities (Year 2003)

Presentations

Predicting Protein-Protein Interactions with Latent Variable Models, Mamitsuka H, Second Asian Joint Workshop on Protein Informatics, Osaka, Japan, 27 February.

Empirical Evaluation of Ensemble Feature Subset Selection Methods for Learning from a High-Dimensional Database in Drug Design, Mamitsuka H, Third IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, USA, 11 March.

Detecting Experimental Noise in Protein-Protein Interactions with Iterative Sampling and Model-based Clustering, Mamitsuka H, Third IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, USA, 12 March.

Efficient Unsupervised Mining from Noisy Data Sets, Mamitsuka H, Third SIAM International Conference on Data Mining, San Francisco, USA, 2 May.

Predicting Protein-Protein Interactions with Hierarchical Latent Variable Models, Mamitsuka H, Third Annual Meeting of Protein Science Society of Japan, Sapporo, Japan, 23 June.

Graph Complexity of Chemical Compounds in Biological Pathways, Yamaguchi A, International Workshop on Bioinformatics and Systems Biology, Dresden, Germany, 18 August.

Hierarchical Latent Knowledge Analysis for Co-occurrence Data, Mamitsuka H, Twentieth International Conference on Machine Learning, Washington DC, USA, 23 August.

Selective Sampling with a Hierarchical Latent Variable Model, Mamitsuka H, Fifth International Symposium on Intelligent Data Analysis, Berlin, Germany, 30 August.

Efficient Mining from Heterogeneous Data Sets for Predicting Protein-Protein Interactions, Mamitsuka H, Fourteenth International Workshop on Database and Expert Systems, Prague, Czech Republic, 3 September.

Current and Future Perspectives of Proteome Informatics, Mamitsuka H, Japan SGI Ltd. Solution Fair 2003, Tokyo, Japan, 5 November.

Finding the Maximum Common Subgraph of a Partial k-Tree and a Graph with a Polynomially Bounded Number of Spanning Trees, Yamaguchi A, 14th Annual International Symposium on Algorithms and Computation, Kyoto, Japan, 15 December.

Efficient Tree-Matching Methods for Accurate Carbohydrate Database Queries, Aoki K, Fourteenth International Conference on Genome Informatics, Tokyo, Japan, 16 December.

Grant

Mamitsuka H, Developing Algorithms for Searching and Finding Small Chemical Compounds Binding to Large Biological Molecules, Grant-in-Aid for Scientific Research on Priority Areas (C), 1 April 2003 - 31 March 2004.

Algorithms and Statistical Analysis of Structural and Functional Properties of Glycans

Glycans, or carbohydrate chains, are "tree"-like structures consisting of monosaccharides and are vital for the development and function of complex multicellular organisms. However, little progress has been made in this field called glycobiology compared to the tremendous progress made in recent years in DNA and proteomics research. This is largely due to the complexity of the biosynthesis and inherent structure of carbohydrate chains. Thus, we have developed a core of algorithms called KEGG Carbohydrate Matcher (KCaM) for querying glycan chains and performed some preliminary statistical analyses that have been promising. KCaM has been incorporated into the KEGG Glycan database.

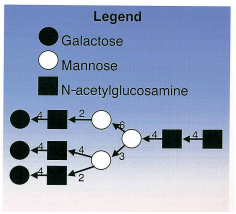

Figure 1. KEGG Glycan ID G00281, an N-linked glycoprotein known to be the primary structure of asialo-carbohydrate units.

Figure 1 is an example of a glycan chain. In graph theoretic terms, each node corresponds to a monosaccharide, and each edge corresponds to a linkage. Note that glycan structures start from the right end, where the root node is located, and branch out toward the left. Each such structure is stored in the KEGG Glycan database, but the usefulness of a database cannot be realized unless it can be queried. This is what prompted the development of KCaM, which is based on existing, known methodologies in theoretical computer science and bioinformatics and consists of two variations; a matching algorithm utilizing dynamic programming and which handles gaps, and one that disallows gaps and finds the largest common subtree based on linkage information.

Specifically, KCaM finds the maximum common subtree of two trees, T_1 and T_2, starting by comparing the leaves of T_1 and T_2 and then moving upward towards the root. This was combined with the Smith-Waterman algorithm, so that each node n_1 from T_1 that forms a match with node m_1 from T_2 stores its match score plus the match information for its children. *Every combination of nodes* is compared in this way, and the subtree with the largest number of matches, which can be found by following the match information at the root and traversing its children down to the leaves, is returned as the final result.

For illustration purposes, Figure 2 compares the glycan of Figure 1, which we call T_1, with a fictional glycan T_2. In this figure, the nodes that are red and purple indicate a match with the red nodes of T_1, but the blue and purple nodes are actually a better match since it corresponds with all of the nodes in T_1. That is, when node n_1 is compared with m_1, traversing up to node m_i, only the red nodes match, but in the pass through the algorithm where node m_i is initially matched with n_1, we end up matching more nodes by the time the comparisons end at the root. Figure 3 illustrates one of the actual resulting matches with the glycan in Figure 1.

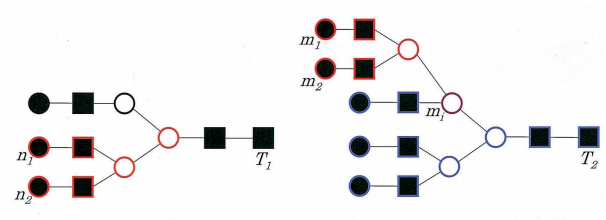

Figure 2. In one pass through the KCaM algorithm where when node n_1 is compared with m_1, traversing up to node m_i, only the red nodes match. However, in the pass where n_1 is matched with m_i, we end up matching more nodes by the time the comparisons end at the root.

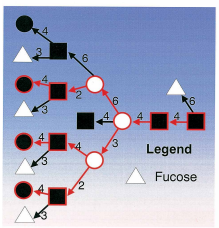

Figure 3. KEGG Glycan ID G05644, with match results with Figure 1 in red. This structure is also an N-linked glycoprotein, published by Yamashita et. al., used in the study of the binding properties of complex-type oligosaccharides on Datura stramonium lectin.

In analyzing the resulting scores from KCaM, the distribution of a random sampling of these scores follows an extreme value distribution, similar to those of sequence similarity scores. We have additionally started incorporating richer information into KCaM in the form of score matrices. Using the scores of matches over the entire KEGG Glycan database, we are developing score matrices taking into account the parent-child relationships in glycan structures. Thus, we can reveal common patterns found in glycans, possibly defining new classes of glycans, and we can statistically estimate the parent-child linkages that frequently appear. Using such tools may assist glycobiologists by providing the clues necessary to glean more detailed information about not only the structure, but also the function of glycans as well.